RedDust: a Large Reusable Dataset of Reddit User Traits

Anna Tigunova, Andrew Yates, Paramita Mirza, and Gerhard Weikum

Max Planck Institute for Informatics, Saarbrücken, Germany
{tigunova, ayates, paramita, weikum}@mpi-inf.mpg.de

Abstract. Social media is a rich source of assertions about personal attributes, such as *I am a doctor* or *my hobby is playing tennis.* Precisely identifying explicit assertions is difficult, though, because of the users' highly varied vocabulary and language expressions. Identifying implicit assertions like *I've been at work treating patients all day* is even more challenging. This paper presents the RedDust data resource consisting of personal attribute labels for over 300k Reddit users across five predicates: profession, hobby, family status, age, and gender. We construct RedDust using a diverse set of high-precision patterns and demonstrate its use as a resource for developing learning models to deal with implicit assertions. RedDust consists of Reddit users ids, the corresponding users' personal attribute labels, and the users' post ids, which may be used to retrieve the posts from a publicly available crawl or from the Reddit API. We discuss the construction of the dataset and show interesting statistics and insights into the data. We also compare different classifiers, which can be learned from RedDust. To the best of our knowledge, RedDust is the first semantic data resource about Reddit users at large scale. We envision further use cases of RedDust for providing background knowledge about user traits, to enhance personalized search and recommendation as well as conversational agents.

1 Introduction

Reddit is a popular social media platform for discussing a wide range of topics. It is an important source of information for data analysis on social media as it provides rich structure, abundance of data and covers a broad range of topics. Reddit is used by approximately 330 million users[1] with 2.8 million comments written each day[2]. Alexa.com ranks it as the 21st most popular website worldwide.

Despite its popular and rich data, few have considered Reddit as a source of data about users' personal traits like their professions, hobbies, and ages. Prior work has focused on Reddit as a source of demographic information, whereas we consider rich predicates like profession and hobbies in addition to demographic ones (age, gender, family status). Such data has many applications,

[1] https://redditblog.com/2018/11/13/holiday-on-reddit/

[2] https://www.digitaltrends.com/social-media/reddit-ads-promoted-posts/

including personalizing healthcare [12], recommendations, search, and conversational agents. To the best of our knowledge, RedDust is the first large scale semantic resource about user traits.

We address this gap by creating a labeled dataset of Reddit users[3] (including their posts and comments) that covers five user attributes: *profession, hobby, family status, age, and gender*. We used three high-precision approaches to identify predicates and their object values in users' posts: *(1)* natural language patterns match assertions like *I am a flight attendant*, *(2)* bracket patterns match structured assertions of users' ages and genders (*I [35m] just broke up with my girlfriend*), and *(3)* flair consists of metadata specific to particular subfora. We use human judgments to validate the high-precision nature of these patterns before performing an analysis of the resulting dataset. We additionally use our data to train several models to label user traits that are expressed implicitly (*I've been fixing sinks all day*) after removing explicit assertions from the data.

This work makes the following contributions:

- We create a dataset of Reddit users traits, which are mined from users' personal assertions with several high-precision techniques. This resource is available at `https://zenodo.org/record/2634977`.
- We perform a thorough analysis of the dataset, which sheds light on its structure and composition.
- We provide a showcase of the dataset usage by performing classification of the labeled attributes with several baseline and state of the art models.

2 Related work

User Profiling in Online Communication: The popularity of social media and online forums brings about massive amounts of user-generated content that is freely accessible. This has opened a great deal of research opportunities on text analysis, in particular to automatically identify latent demographic features of online users for personalized downstream applications such as personalized search or recommendation. Such latent demographic attributes include age and gender [1,2,3,5,7,13,21,22,24,30], personality [11,24], regional origin [5,21], political orientation and ethnicity [15,19,20,21,30], as well as occupational class that can be mapped to income [8,18].

Most prior works on automatically identifying users' latent attributes from online communication rely on classification over hand-crafted features such as word/character n-grams [1,3,21], Linguistic Inquiry and Word Count (LIWC) [16] categories [11,19,20], topic distributions [7,15,18] and sentiment/emotion labels of words derived from existing emotion lexicon [11,15,19,20]. The recently prominent neural network approaches have also been adopted to solve the task [2,13,28,30]. The vast majority of these works also rely on features specific to social media such as emojis and hashtags [3,15,21], users' profile descriptions [3,15,30] and communication behavior [15,21], and social network structure

[3] `https://zenodo.org/record/2634977`

[13,15,30], with only [1,2,18,28] inferring users' latent attributes based solely on user-generated text.

Dataset for User Profiling: Automatic methods, particularly supervised learning approaches, for identifying users' personal attributes require a collection of user-generated content labelled with personal attributes of interest. Most of existing works mentioned above focus on user-generated content from Twitter, with a few exceptions that explore Facebook [22,24] or Reddit [5,6,11] posts.

Data collection was mostly done via: manual annotation after a focused search with specific keywords or hashtags [18,21], public profile linked to Twitter profile description [3,7], self-reports as part of an online survey [6,7,19,20,22,25], or pattern-based extraction approach (e.g., `(I|i) (am|'m|was) born in + number (1920-2013)`) on user profile description or user posts [5,13,26,28]. Several works [1,2] made use of labelled datasets published within the shared task on *author profiling* organized by the CLEF PAN lab [9,17].

There have been less effort on identifying demographic attributes of Reddit users compared with the body of work exists for Twitter users, although Reddit posts have been exploited for other purposes such as determining users' personality [11], mental health condition [4], domestic abuse [23] and irony detection [31], among others. In [27], the authors investigate how the topic of subreddit influences the gender ratio within it. The study was performed on 100 subreddits grouped by interest, gender information about the users were collected by guessing it from their usernames, which is arguably a low-precision strategy. Smaller scale Reddit datasets exist for gender, age and location attributes [5,6], which are unfortunately not publicly available. As far as we know, we are the first to consider *hobbies* as a personal attribute of interest to be identified from online communication.

3 Background

Reddit[4] is a social news website and forum where registered members can submit content including links, text posts, and images, which are then voted up or down by other members. Before elaborating on the creation of our dataset derived from Reddit posts, we describe several concepts on Reddit that are relevant for the data collection process.

Posts and Comments: Discussions on Reddit are organized in threads, which are initiated by an original *post* and may contain *comments* replying to the post and to other comment. This creates a hierarchical structure that resembles a conversation between users. Both posts and comments can be a textual content, a link with anchor text or images.

Subreddits: Reddit is organized into subreddits, which are fora that focus on specific topics. Those can be split by interest (sports, politics, etc), by country or

[4] `https://www.reddit.com/`

community, type of content (text, gifs, videos), and so on. Subreddits have their own rules, but any registered user can create them. By convention, subreddits are prefixed with `/r`. For example, users discuss hockey in the `/r/hockey` subreddit.

Flairs: Flair is user or post metadata that is a unique feature of Reddit. Flair is generally a small image with a short text description that is attached to a post or a username. Flairs can be defined differently for specific purposes by each subreddit. For example, in `/r/travel` subreddit they may indicate the *country* of the user, *gender* in `/r/AskMen` and `/r/AskWomen` or users' *favorite teams* in `/r/hockey`. Flairs for posts can be useful to filter and search for a particular content.

4 Dataset Creation

RedDust is a dataset containing a collection of Reddit users, or *redditors*. Each redditor in the dataset is associated with posts and comments they produce, and personal attributes resulting from pattern-based mining approach as well as flairs. In this work, we consider five personal attributes including *gender*, *age*, *family status*, *profession* and *hobby*. The dataset is created from the openly published Reddit dump[5], which spans between 2006 and 2018.

There are several criteria on which users and posts/comments that are included in RedDust, i.e., users who posted between 10 and 100 posts/comments, and posts/comments containing between 20 and 100 terms after filtering. We filtered out hyperlinks and user mentions (i.e., `@nickname`) from the original posts/comments.

Some subreddits are likely to contain many false positives, such as those concerned with video games or role playing. This leads to personal assertions talking about the users' projected persona in a particular context (e.g., *I am a priest looking for a guild*). To mitigate this source of false positives, we blacklisted subreddits about gaming, fantasy, and virtual reality from the top 500 subreddits sorted by number of unique users. Posts made to blacklisted subreddits were discarded. The list of blacklisted subreddits can be found in Appendix 2. Similarly, we do discard posts that contain quotations in order to reduce the possibility of the user referring to a third person (*... and he shouted "Hands on the counter, I am a cop!"...*).

For attributes that usually have a unique value (i.e., *gender*, *age* and *family status*) we also exclude users who state multiple different values to avoid introducing false positives. For *profession* and *hobby*, we allow each user to have multiple attribute values. The age of a given user is calculated relative to his or her age when writing the most recent comment.

In the following subsections we discuss the particular techniques used to extract object values for each predicate.

[5] `https://files.pushshift.io/reddit/`

4.1 Gender

In author profiling work, gender has been the most popular user attribute to predict [5,27,29]. In our dataset we consider gender as a binary predicate (male or female) and discard other values as in prior work.

We look for self-reported gender assertions, which provides a high-precision source of labels. User names are not considered as a means for gender classification (as done in [27]), because this does not achieve high precision. We identified object values for the gender predicate using the following methods:

- **Natural language patterns**
  Following [5] we manually created a set of patterns that indicate a specific gender. They have the general form of `(I am|I'm) a? <gender word>`, meaning that matches should contain *I am* or *I'm*, optionally followed by *a*, and then a word that indicates the author's gender like *man* or *mother*. A comprehensive list of the patterns we used is given in Table 1, and the indicative gender words are shown in Table 2. Although the user's gender can be expressed in a longer snippet like *I am a great mother*, we don't allow extra words like *great* to appear before the gender word. This reduces false positives from statements like *I'm a far cry from a mother*.
  Still, there is a number false positive patterns, which are tricky to avoid. Those could be imaginary situations (*I dreamed I am a mother*), reported speech (*she said "I am a mother"* - we don't consider the sentences with quotes for this purpose) and some other (*I am a woman lover*). Those are hard to avoid automatically and thus the manual examination over the extracted posts should be run.
- **Bracket patterns** In situations where the user wants to indicate their demographic information, posts contain direct indicators of the user's age and gender of the user (*I [35m] went hiking*). This is common in relationship-related subreddits where the user's age and gender are often relevant to the topic of discussion. These cues are generally written in round or square brackets. To reduce false positives, we do not consider such patterns when they appear without brackets. To capture gender and age expressed in this way, we look for patterns of the form `(I|I'm|me) <number>(m|f)`.
- **Flair**
  Similar to [29], we consider gender-indicating flair attached to users. This logic is subreddit-specific, so we restrict ourselves to common subreddits. For example, for the subreddits `/r/AskWomen` and `/r/AskMen` the flair is one of (*male, female, trans*, etc), whereas in `/r/tall` and `/r/short` the flair is (*pink, blue, other*).

4.2 Age

We label users' posts with age predicate using the same techniques as with gender:

- **Patterns** For the age predicate we have considered the following patterns: (i) `I (was|am) born in <four digit year>`; (ii) `I (was|am) born in <two digit year>s`; (iii) `I was born on <day, month, year>`; (iv) `I am <number> years old`. After that we calculated the age for patterns (i)-(iii) by subtracting the birth year from the post's publishing date. In addition to those we consider a special case *I am* + integer, which indicates the age, only when it is followed by punctuation or conjunction (*and*, *but*, ...). This is important to avoid false positives (*I am 6 feet tall*).
- **Bracket patterns** In this case age was jointly collected with gender, as described in the previous section.

Finally we checked that all obtained ages are within the range between 10 and 100 years old, since users under 13 are not allowed to register and there are unlikely to be many users above 100 years old. This is helpful for reducing false positives, such as those in conditional sentences (*as if I am 5 years old*).

4.3 Family status

We consider family status as a binary predicate indicating whether a person is single or has a partner. Similar to the gender predicate we collected identifier words. We distinguished two cases:

- The speaker refers to himself: the post should match the pattern `I am <self status indicator>`, where `<self status indicator>` is a word like *married* or *single*.
- The speaker refers to his partner: the pattern is defined as `(my|I have a) <partner status indicator>`, where `<partner status indicator>` is a word like *wife* or *spouse*.

We additionally collected negative matches (*I am not..*, *I don't have..* for the opposite value of the predicate to increase the volume of labeled data. Furthermore, among the indicative words we did not consider the word *single* alone, as it often gives false matches (*single player*), therefore, for this word in particular we made a restriction that it should be followed by punctuation, *and*, *or*, or should be in the form like *single mother*.

4.4 Profession

To obtain profession labels we took a list of occupation names from Wikipedia, by recursively adding all subcategories starting from `https://en.wikipedia.org/wiki/Lists_of_occupations`. The resulting list we obtained consisted of about a thousand professions and contains a lot of fine grained occupations, some of which were redundant or ambiguous. Our strategy is to capture as many possible profession assertions as possible, giving users of RedDust the opportunity to filter and group the professions depending on their specific use cases.

We used this list of profession names to construct patterns of the form `(I am|I'm) a <profession name>`. We did not use other approaches, such as flair, because they are less common and less reliable with this predicate. After performing pattern matching, we were left with 832 unique professions in the dataset.

attribute	**pattern**
profession	(I am|I'm) a <profession name>
hobby	<phrase> (*I enjoy*, ...) <hobby> (*books*, *sports*...)
gender	(I am|I'm) a? + <gender word (*lady*, *father*, etc.)>
age	I (was|am) born in <four digit year> I (was|am) born in <two digit year>s I was born on <day, month, year> I am <number> years old
family status	I am <self status indicator> (*married*, *divorced*, *single*, ...) (my|I have a) <partner status indicator> (*wife*, *boyfriend*, ...)

Table 1: Patterns for extracting ground truth labels from Reddit posts.

predicate	keywords and key phrases
gender	female: *woman*, *female*, *girl*, *lady*, *wife*, *mother*, *sister* male: *man*, *male*, *boy*, *husband*, *father*, *brother*
family status	single: *single*, *divorsed*, *widow*, *spouseless*, *celibate*, *unwed*, *fancy-free* married: *married*, *engaged*, *dating*, *boyfriend*, *spouse*, *girlfriend*, *fiancee*, *lover*, *partner*, *wife*, *husband*
hobby	*I am fond of*, *I'm fond of*, *I am keen on*, *I'm keen on*, *I like*, *I enjoy*, *I go in for*, *my hobby is, I take joy in, I adore, I love, I play, I fancy, I am a fan of, I'm a fan of, I am fascinated by, I'm fascinated by, I am interested in, I'm interested in, I appreciate, I practise I am mad about, I'm mad about*

Table 2: Predicate keywords and key phrases.

4.5 Hobby

Similar to the profession predicate, we obtained a list of Hobbies from Wikipedia[6] and searched for mentions of hobbies in this list. We used a diverse set of patterns of the form `<phrase> <hobby>`, where `<phrase>` is a phrase like *my hobby is* or *I enjoy*. The full list of phrases is shown in Table 2. After performing pattern matching, we were left with 336 unique hobbies in the dataset.

	gender	age	family status	profession	hobby
precision	0.96	1.0	0.86	0.96	0.94
number of false positives	2	0	7	2	3
number of disagreements	2	2	8	2	9

Table 3: Inter-rater agreement and false positive numbers for manual evaluation.

[6] https://en.wikipedia.org/wiki/List_of_hobbies

4.6 Labeling evaluation

To validate the high-precision nature of our labeling approach, we asked three human annotators to verify the correctness of labels for each predicate. To do so, we randomly sampled 50 labeled posts for each predicate and asked annotators to indicate whether the given label matched the user's actual assertion. The decision to accept or reject the label was based on a majority vote from the annotators.

The results of this evaluation are shown in Table 3. In total there were 23 instances without perfect annotator agreement (out of 250 total instances), which indicated 14 false positives after taking a majority vote. Half of these false positives came from the family status predicate, due to ambiguous usage of words like *single* and *lover* in phrases like *I am a lover of the game*. Despite these false positives, the minimum precision for any predicate considered was 86%, and four out of the five predicates had a precision of at least 94%.

4.7 Privacy

We note that our dataset consists of predicate-object pairs for real subjects (Reddit users) who may desire to edit or delete their posts. We take several steps in order to protect users' privacy and to preserve their ability to remove their posts. First, we provide only post ids, so that users may opt to delete their posts and thus prevent them from being included in copies of the dataset. Second, we do not disclose any specific usernames, so that users retain the ability to sever links between their posts (e.g., by deleting a post's author information but leaving the post text[7]).

5 Data statistics and analysis

In this section we present the quantitative and qualitative analysis of the RedDust dataset. In Table 4 we present the overall numerical statistics of the dataset. For the counts of the particular attribute values we refer to the Appendix 1.

In Figure 1 we plot the chart of the number of users per each count of posts within the admissible 20-100. From this plot we conclude that the users in our dataset tend to have a small number of posts

	gender	age	family status	profession	hobby	total
Num of users	54,879k	122,198k	11,770k	74,861k	89,068k	352,776k
Num of posts	2,488M	5,801M	0,558M	3,629M	4,422M	16,901M
Num of subreddits	28,246k	44,070k	14,760k	37,485k	41,314k	165,875k

Table 4: Overall dataset statistics

[7] This is supported by Reddit.

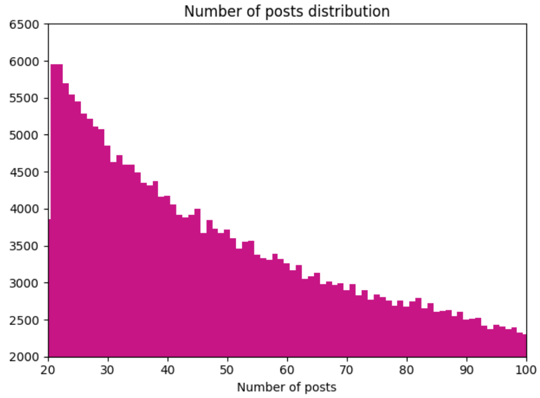

Fig. 1: Counts of users having X posts

As mentioned before, the profession attribute has 832 possible values in total. However, just about 5% of professions have over 2000 users, which we consider to be a sufficient number.

5.1 Analysis of multiple predicates

In our dataset almost 19k users have two predicate values known, 980 have three and 28 have four. From that one can observe that about 6% of the users have labels for multiple predicates. That is, however, sufficient enough to be used to train multi-label prediction models.

For those users who have several attributes known it is interesting to look at interplay between different traits. As one example of that we considered a pair of profession and hobby predicates (as they have the greatest number of values). In Figure 2 we plotted a heat map which represents the co-occurence of the values for these two predicates. Here and in the following experiments we limited the number of professions and hobbies to the top ones, sorted by the number of users.

There we can observe quite obvious patterns as guitar being most practised by drummers and musicians in general; runners have running as the main interest; students are interested in video games 5 times as much as any other profession; and curiously shooting is popular among photographers, probably because of *shooting* being an ambiguous term.

We also considered other pairs of predicates. In Figure 3 we show the gender distribution for certain professions. Here we can reveal some sexist patterns (like *nanny* being mostly feminine and *programmer* masculine) as well as some surprising insights (prevalence of female *runners* and *bartenders*).

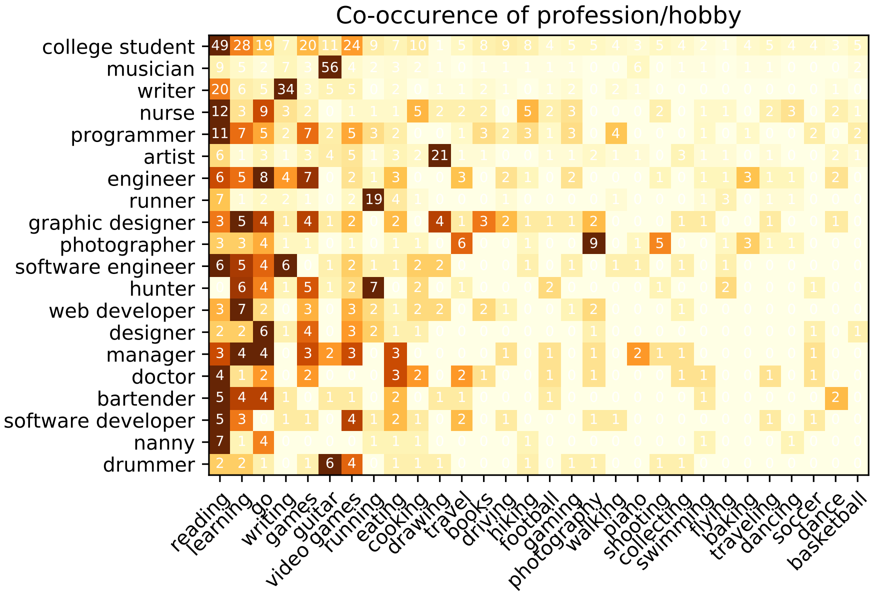

Fig. 2: Co-occurrence of the most common professions and hobbies.

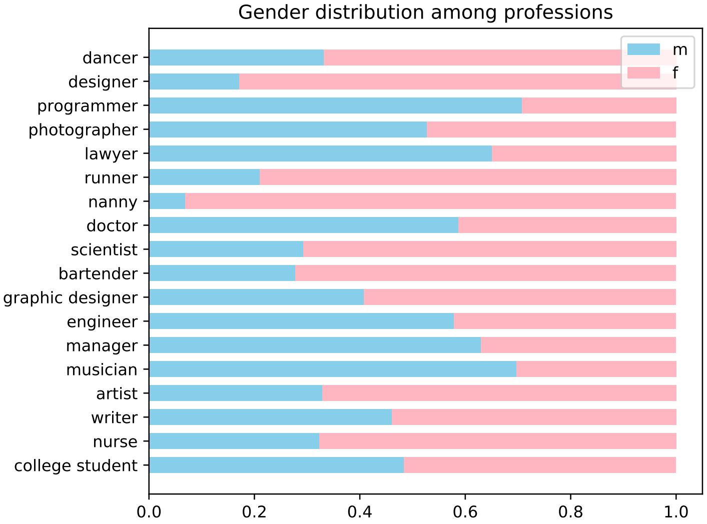

Fig. 3: Gender distribution among professions.

5.2 Subreddit analysis

Together with user labels and posts we can also get the subreddits these posts were written in. We found it also useful to investigate the dependencies between predicates and subreddits. For instance, Figure 4 shows both the distribution of genders in out dataset as well as the distribution of most popular subreddits for each gender.

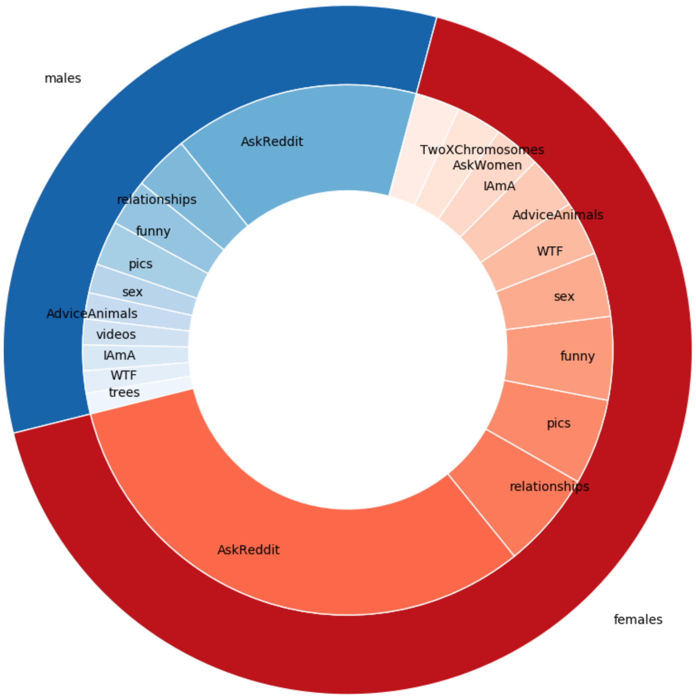

Fig. 4: Most popular subreddits for genders

From this chart it can be seen that females are more common in our dataset, which does not match the statistics from the other resources [8]. However, our dataset does not give the true outlook on Reddit demoography and the gender ditribution that we got may be contributed to (i) females being more self-disclosing (ii) peculiarities of our labeling patterns.

In addition to that, we can observe that the list of top subreddits for both genders is almost the same, including even the order. However, these are the most popular subreddits in general, therefore, they are supposed to be at the top of he list for any value for all attributes; and to get a deeper insight we might like to remove them to find the most typical subreddits for values of a specific predicate.

[8] `https://www.techjunkie.com/demographics-reddit/`

We did this kind of analysis for the profession predicate. We defined the most common subreddits to be the ones in which at least 15 users from our dataset have posted and removed them from consideration. After that for each profession P we calculated the scores for each resulting subreddit as $Pr(sub|prof = P)$. We rank then the subreddits by probability and present top results for some professions in Table 5.

profession	typical subreddits
nurse	*nursing, MakeupAddiction, loseit, AskWomen, StudentNurse*
photographer	*photography, photocritique, itookapicture, Instagram, analog*
software developer	*programming, cscareerquestions, learnprogramming, webdev,* Entrepreneur
web developer	*webdev, web_design, forhire, Entrepreneur, programming*

Table 5: Typical subreddits for profession

We can see that the obtained top subreddits are one the one hand very relevant to the particular profession specification, on the other hand they are also not likely to occur in other profession sets.

5.3 Word analysis

Finally, we analyze the vocabulary that distinguished people of different predicate values. For that we counted and sorted the number of word occurrences for each value. We remove the words that are too common (a word encountered in speech of more than 15 values) and too rare (shall be at least used by 4 values). In Table 6 we list the resulting typical words for some professions, sorted by the number of occurrence.

profession	typical words
writer	*wordpress, spacebar, erotica, screenwriter, goblins, scandalous*
bartender	*bartend, limes, bitters, tippers, dreamcast, cognac*
musician	*clifford, percussion, composing, triangles, zack, concerto, artist*
web developer	*iamas, fonts, stackoverflow, serif, sharply, lynda, freelancing*
lawyer	*statute, litigation, counsel, prosecutor, plea, fafsa, defendant*

Table 6: Typical words per profession

6 Experiments

To show one of the possible usecases of RedDust we run several classifiers on our dataset to predict correct attributes of the users. We performed binary classification for gender and family status and the multi-class classification for the rest of the predicates. For each attribute we form training and test sets by splitting the initial dataset in a 9:1 proportion. Each training instance for the classifiers is a user represented by all his posts, each of which is a bag of words. Following prior work [28], we remove punctuation, stopwords, user names, hash tags and hyperlinks from the posts. Input words are represented using word2vec embeddings trained on GoogleNews. [14] Since our goal is to label users based on only their implicit assertions, we remove all posts which explicitly mention attribute values (i.e., the posts that were used for labeling users). We evaluate the following classification methods:

Logistic regression. We used a multinomial logistic regression classifier to get the probabilities for the values of each attribute. The input to the classifier is the average of embeddings of all words in all posts of the user.

Multi-layer perceptron (MLP). We applied a shallow MLP classifier with one hidden layer of size 100 and ReLU activation, which was trained for 200 epochs. The input to MLP is the same as for logistic regression.

Convolutional neural network. The input to the CNN model is the concatenation of all words for a given user. The model uses filters of sizes 2 and 3 to produce feature maps, which are then concatenated and classified with a fully-connected layer. We used 64 convolutional filters and trained for 100 epochs.

Hidden Attribute Models [28]. We selected $\text{HAM}_{\text{CNN-attn}}$, a model that utilizes a CNN with attention, which was shown to perform the best on classifying personal attributes. This model considers the posts of the users in a hierarchical way: first, the representations of words are put through a CNN for each post separately to create the post's latent representation. Then the model applies attention mechanism to them, to find the importance weights of each post. After that the weighted average of the posts is taken and finally this weighted average is put through a fully-connected layer to get the probabilities for each class. We used the following hyperparameters for $\text{HAM}_{\text{CNN-attn}}$: $numberoffilters = 128, hiddensizeofattentionlayer = 150, numberofiterations = 70$.

6.1 Binary predicates

To evaluate the classifiers on binary predicates we compute accuracy and area under the curve (AUROC) metrics. AUROC is computed by varying scores threshold and measuring the true positive rate and false positive rate for the model's predictions. The results of evaluation on binary predicates are in Table 7.

6.2 Multi-label predicates

For the multi-label predicates hobbies and professions, we select a subset of object values that have a sufficient number of users. To do so, we started with the the values associated with more than 250 users. We then manually cleaned these values by merging similar ones (e.g., *police officer* and *cop*). This left us

	gender		family status	
	AUROC	acc	AUROC	acc
logreg	0,49	0,57	0,5	0,5
MLP	0,4891	0,57	0,63	0,63
CNN	0,88	0,8	0,9	0,81
HAM	0,91	0,86	0,89	0,82

Table 7: Results of evaluation of classifiers on the binary predicates.

with 69 unique professions and 89 unique hobbies to use in our evaluation. This list can be found in Appendix 3.

For all multi-label predicates (age, profession and hobby) we report performance in terms of AUROC and two ranking metrics: mean first relevant (MFR) and normalized discounted cumulative gain (nDCG). To compute AUROC in multi-label case we binarize the labels and compute one-vs-all scores. To get the models' ranking of answer we rank the predicted probabilities for each object value. We compute MFR, a variation on MRR (mean reciprocal rank), as an average of the rankings of all instances in the test set. We prefer MFR over MRR since recent work has identified issues with MRR's formulation. [10] We use only two labels with nDCG: 1 for a correct object value and 0 for an incorrect one.

	age			profession			hobby		
	MFR	AUROC	nDCG	MFR	AUROC	nDCG	MFR	AUROC	nDCG
logreg	3	0,78	0,9	34,6	0,74	0,34	39,05	0,79	0,44
MLP	3	0,82	0,96	32,95	0,78	0,35	37,3	0,8	0,46
CNN	2,78	0,85	0,95	12,7	0,83	0,45	18,6	0,77	0,3
HAM	2,44	0,88	0,96	14,5	0,85	0,47	18,3	0,8	0,3

Table 8: Results of evaluation of the classifiers on multi-label predicates.

The results of classification for the multi-labeled case are presented in Table 8. As one can observe from both Tables 7 and 8, perform reasonably well but cannot accurately predict the correct object value for the more difficult predicates. In addition to the fact that this is simply a hard classification task that is underexplored, there are two clear factors that hinder models' performance: (i) the distribution of labels is very skewed (for instance, the great prevalence of teenagers and students), and (ii) the groupings of some professions and hobbies are subjective and may be ambiguous. Thus we conclude that there is a need for further research on identifying implictly-stated attribute values. We hope our dataset will facilitate the development of such models, as well as other researching on user traits.

7 Conclusion

In this work we described RedDust, a large semantic resource about Reddit users' personal traits that was created using a combination of high-precision patterns and Reddit metadata. RedDust consists of over 350,000 users labeled with object values for five predicates. We demonstrate one use case for this data by training models to predict users' traits based on only implicit assertions. RedDust is available at `https://zenodo.org/record/2634977`.

References

1. Basile, A., Dwyer, G., Medvedeva, M., Rawee, J., Haagsma, H., Nissim, M.: N-GrAM: New Groningen Author-profiling Model—Notebook for PAN at CLEF 2017. In: Working Notes Papers of the CLEF 2017 Evaluation Labs (2017)
2. Bayot, R.K., Gonçalves, T.: Age and gender classification of tweets using convolutional neural networks. In: Machine Learning, Optimization, and Big Data. Springer International Publishing, Cham (2018)
3. Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on twitter. In: Proceedings of EMNLP'11 (July 2011)
4. Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., Goharian, N.: ”SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions”. In: Proceedings of the 27th International Conference on Computational Linguistics (2018)
5. Fabian, B., Baumann, A., Keil, M.: Privacy on reddit? towards large-scale user classification. In: Proceedings of ECIS'15 (2015)
6. Finlay, S.C.: Age and gender in reddit commenting and success (2014)
7. Flekova, L., Carpenter, J., Giorgi, S., Ungar, L., Preoţiuc-Pietro, D.: Analyzing biases in human perception of user age and gender from text. In: Proceedings of ACL'16 (Volume 1: Long Papers) (2016)
8. Flekova, L., Preoţiuc-Pietro, D., Ungar, L.: Exploring stylistic variation with age and income on twitter. In: Proceedings of ACL'16 (Volume 2: Short Papers) (2016)
9. Francisco Manuel, Rangel Pardo, Rosso, P., Potthast, M., Stein, B.: Overview of the 5th Author Profiling Task at PAN 2017: Gender and Language Variety Identification in Twitter. In: Working Notes Papers of the CLEF 2017 Evaluation Labs (2017)
10. Fuhr, N.: Some common mistakes in ir evaluation, and how they can be avoided. In: ACM SIGIR Forum. vol. 51, pp. 32–41. ACM (2018)
11. Gjurković, M., Šnajder, J.: Reddit: A gold mine for personality prediction. In: Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media, NAACL-HLT'18 (2018)
12. Gyrard, A., Gaur, M., Shekarpour, S., Thirunarayan, K., Sheth, A.: Personalized health knowledge graph. In: First International Workshop on Contextualized Knowledge Graphs (2018)
13. Kim, S.M., Xu, Q., Qu, L., Wan, S., Paris, C.: Demographic inference on twitter using recursive neural networks. In: Proceedings of ACL'17 (Volume 2: Short Papers) (July 2017)
14. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: Advances in Neural Information Processing Systems. pp. 3111–3119 (2013)

15. Pennacchiotti, M., Popescu, A.M.: A machine learning approach to twitter user classification. In: Proceedings of ICWSM'11 (2011)
16. Pennebaker, J.W., Francis, M.E., Booth, R.J.: Linguistic inquiry and word count: Liwc 2001. Mahway: Lawrence Erlbaum Associates **71** (2001)
17. Potthast, M., Rangel, F., Tschuggnall, M., Stamatatos, E., Rosso, P., Stein, B.: Overview of PAN'17: Author Identification, Author Profiling, and Author Obfuscation. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 7th International Conference of the CLEF Initiative (CLEF 17) (2017)
18. Preoţiuc-Pietro, D., Lampos, V., Aletras, N.: An analysis of the user occupational class through twitter content. In: Proceedings of ACL/IJCNLP'15 (Volume 1: Long Papers) (July 2015)
19. Preoţiuc-Pietro, D., Liu, Y., Hopkins, D., Ungar, L.: Beyond binary labels: Political ideology prediction of twitter users. In: Proceedings of ACL'17 (Volume 1: Long Papers) (2017)
20. Preoţiuc-Pietro, D., Ungar, L.: User-level race and ethnicity predictors from twitter text. In: Proceedings of COLING'18 (2018)
21. Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in twitter. In: Proceedings of SMUC'10 (2010)
22. Sap, M., Park, G., Eichstaedt, J., Kern, M., Stillwell, D., Kosinski, M., Ungar, L., Schwartz, H.A.: Developing age and gender predictive lexica over social media. In: Proceedings EMNLP'14 (October 2014)
23. Schrading, N., Alm, C.O., Ptucha, R., Homan, C.: An analysis of domestic abuse discourse on reddit. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 2577–2583 (2015)
24. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E.P., Ungar, L.H.: Personality, gender, and age in the language of social media: The open-vocabulary approach. In: PloS one (2013)
25. Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M., Shah, A., Kosinski, M., Stillwell, D., Seligman, M.E., et al.: Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one **8**(9), e73791 (2013)
26. Sloan, L., Morgan, J., Burnap, P., Williams, M.: Who tweets? deriving the demographic characteristics of age, occupation and social class from twitter user meta-data. PloS one **10**(3), e0115545 (2015)
27. Thelwall, M., Stuart, E.: She's reddit: A source of statistically significant gendered interest information? Information Processing & Management (2018)
28. Tigunova, A., Yates, A., Mirza, P., Weikum, G.: Listening between the lines: Learning personal attributes from conversations. In: The Web Conference. ACM (2019)
29. Vasilev, E.: Inferring gender of reddit users
30. Vijayaraghavan, P., Vosoughi, S., Roy, D.: Twitter demographic classification using deep multi-modal multi-task learning. In: Proceedings of ACL'17 (Volume 2: Short Papers) (July 2017)
31. Wallace, B.C., Kertz, L., Charniak, E., et al.: Humans require context to infer ironic intent (so computers probably do, too). In: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). vol. 2, pp. 512–516 (2014)